The Federation of Earth Science Information Partners

SCIENCE DATA ANALYSIS ON THE CLOUD

ESIP Cloud Computing Cluster

Thomas Huang and Phil Yang

Agenda

- Invited speakers
  - **Petr Votava**, NASA Earth Exchange (NEX): Early Observations on Community Engagement in the Cloud
  - **Wes Lloy**, The Virtual Machine Scaler: Infrastructure Management Support for Scientific Modeling on IaaS Clouds
- Eucalyptus 4 Upgrade
- Cloud Computing Cluster onward

NASA Earth Exchange (NEX): Early Observations on Community Engagement in the Cloud

Petr Votava
NASA Ames Research Center

Abstract: NASA Earth Exchange (NEX) is a collaborative platform that combines state-of-the-art supercomputing, Earth system modeling, remote sensing data from NASA and other agencies, and a scientific social network to provide an environment in which users can explore and analyze large Earth science data sets, run modeling and analysis codes, collaborate on new or existing projects, and share results within and/or among communities. Through the deployment of virtualization technologies, an opportunity exists to create complete modeling and analysis environments that are customizable, “archiveable” and transferable. Allowing users to instantiate such environments on large compute infrastructures that are directly connected to data archives may significantly reduce costs and time associated with scientific efforts by alleviating users from redundantly retrieving and integrating data sets and building codes as well as provides a mechanism for sharing their work with the community. NEX is pursuing this development through OpenNEX partnership with Amazon, Inc. as well as locally through the NEX OpenSandbox, which provides private cloud environment collocated with NASA supercomputing center. This talk will focus on some of the reasons for pursuing the cloud as one of our platforms as well as our first observations on trying to generate interest and participation from a large community of geoscientists, software engineers, students and the general public.

The Virtual Machine Scaler: Infrastructure Management Support for Scientific Modeling on IaaS Clouds

Wes Lloyd

Colorado State University

Abstract: (IaaS) clouds provide a new medium for deployment of scientific modeling applications. Harnessing advancements in virtualization, IaaS clouds can provide dynamic scalable infrastructure to better support the computational demands of modeling. Providing scientific modeling "as-a-service" requires dynamic scaling of server infrastructure to adapt to changing user workloads. This paper presents the Virtual Machine (VM) Scaler, an autonomic resource manager for IaaS Clouds. We have developed VM-Scaler, a REST/JSON-based web services application which supports infrastructure provisioning and management to support scientific modeling for the Cloud Services Innovation Platform (CSIP). VM-Scaler harnesses the Amazon Elastic Compute Cloud (EC2) application-programming interface to support model-service scalability, cloud management, and infrastructure configuration for supporting modeling workloads. VM-Scaler provides "cloud control" while abstracting the underlying IaaS cloud from the end user. VM-Scaler is extensible to support any EC2 compatible cloud and currently supports the Amazon public cloud and Eucalyptus private clouds versions 3.x. VM-Scaler provides a platform to improve scientific model deployment by supporting experimentation with: hot spot detection schemes, VM management and placement approaches, workload profiling techniques, and model job scheduling/proxy services. VM-Scaler has been exercised to scale model services and execute large workloads in the cloud for several research and operational environmental science models used by the US Department of Agriculture.

Session 3041: Leveraging Enabling Technologies and Architectures to enable Data Intensive Science

The objective of this session is to share innovative concepts, emerging solutions, and applications for Big Earth and Space Data to enable Data-Intensive Science. Data-Intensive Science defines three high-level activities: capture, curation, and analysis of data. Being able to handle massive amount of data impacts our architectural decisions and approaches. Topics include demonstration, studies, methods, and/or architectural discussion on

- Common enabling technologies
- Automated techniques for data analysis
- Scienhttp://sites.agu.org/ce analysis and visualization
- Real time decision support
- Implication of Data Intensive science to education
- Data management lifecycle functions from data capture through analysis
- Architecture that spans multiple data systems and organizations

Conveners:

- Thomas Huang, Jet Propulsion Laboratory
- Rahul Ramachandran, NASA Global Hydrology Resource Center (GHRC)
- Dan J. Crichton, Jet Propulsion Laboratory
- Morris Riedel, Juelich Supercomputing Center, Julich, Germany

AGU Fall Meeting

Abstracts Due: August 6, 2014

Session 1832: Cloud Computing for Geosciences

Geosciences pose cyber-infrastructure and computing challenges from disruptive hazards events, complex model configuration, interoperability in data/service sharing, and many others. Cloud Computing emerged as one of the most promising solutions. This session welcomes contributions that demonstrate addressing these challenges with cloud-computing. Topics include but not limited to cloud computing related research and development for

- data management
- interoperability
- computing support
- real time decision support
- knowledge management
- model configurations and simulations
- visual analytics, and
- cloud-based uncertainty quantification and sensitivity analyses.

Conveners:

- Phil Yang, George Mason University
- Thomas Huang, Jet Propulsion Laboratory
- Alex Sun: University of Texas
- Wenming Ye, Microsoft Corporation

NASA ESDSWG – Data-Intensive Architecture (DIA)

- How to scale our system architecture so it can respond to an exponential increase in the amount of data we are serving?
- This working group focuses on this data-intensive reality in an effort to deliver recommendations, reference architectures and eventually a roadmap to better serve our current and future missions, and our science communities.
- Some of the data intensive challenges are:
  - Lifecycle: Capture, Curate, Analyze
  - Data Movement
  - Data Coordination
  - Data Reduction
  - Topography Optimization
  - Decision Support
- Submit Your Use Cases: **`http://bit.ly/1rUq3bj`**

ESIP Cloud Computing Cluster onward

THANKS